Statistical Modelling 2005; **5**: 1–19

Random effect models for repeated measures of zero-inflated count data

Yongyi Min[1] **and Alan Agresti**[2]
[1]Statistical Division, The United Nations, New York, USA
[2]Department of Statistics, University of Florida, Florida, USA

Abstract: For count responses, the situation of excess zeros (relative to what standard models allow) often occurs in biomedical and sociological applications. Modeling repeated measures of zero-inflated count data presents special challenges. This is because in addition to the problem of extra zeros, the correlation between measurements upon the same subject at different occasions needs to be taken into account. This article discusses random effect models for repeated measurements on this type of response variable. A useful model is the hurdle model with random effects, which separately handles the zero observations and the positive counts. In maximum likelihood model fitting, we consider both a normal distribution and a nonparametric approach for the random effects. A special case of the hurdle model can be used to test for zero inflation. Random effects can also be introduced in a zero-inflated Poisson or negative binomial model, but such a model may encounter fitting problems if there is zero deflation at any settings of the explanatory variables. A simple alternative approach adapts the cumulative logit model with random effects, which has a single set of parameters for describing effects. We illustrate the proposed methods with examples.

Key words: cumulative logit model; generalized linear mixed model; hurdle model; negative binomial model; nonparametric mixture model; zero-inflated Poisson model

Data and software link available from: http://stat.uibk.ac.at/SMIJ

1 Introduction

We consider models for count responses with excess zeros relative to what standard distributional assumptions, such as the Poisson, can predict. In the literature, ‘zero-inflated count data’ refers to data for which a generalized linear model has lack of fit due to disproportionately many zeros. Such data are common in many applications, especially when many subjects have zero observations, yet many also have much larger observations so that the overall mean need not be near zero. An example of a variable that one might expect to be zero inflated is the number of times a subject used medical services in the previous year: some subjects may have a zero observation because of chance, whereas others may have a zero observation because of a ‘doctor avoidance’ phobia.

Address for correspondence: Yongyi Min, Statistical Division, The United Nations, 2 UN Plaza, DC2-1404, NY 10017, USA. Tel: +1 212 963 9296; E-mail: min3@un.org

10.1191/1471082X05st084oa

There is considerable literature on modeling cross-sectional zero-inflated count data, using the hurdle model (Arulampalam and Booth, 1997; Mullahy, 1986) and the zero-inflated count model (Lambert, 1992; Shankar *et al.*, 1997). The hurdle model is a two-part model for count data. One part is a binary model for whether the response outcome is zero or positive. If the outcome is positive, the 'hurdle is crossed'. Conditional on a positive outcome, the second part uses a truncated model that modifies an ordinary distribution by conditioning on a positive outcome. For instance, this might be a truncated Poisson distribution or a truncated negative binomial distribution. The hurdle model can handle both zero inflation and zero deflation.

A separate strand of the literature pertains solely to zero inflation. With this approach, two types of zeros can occur: one comes from the zero state and the other from the ordinary count distribution state. That is, the relevant distribution is a mixture of an ordinary count model, such as the Poisson or negative binomial, with one that is degenerate at zero (Lambert, 1992). Such zero-inflated count models are more natural than a hurdle model when it is reasonable to think of the population as a mixture, with one set of subjects that will have only a zero response and other subjects that may have a zero response, such as the use of medical services example mentioned earlier.

Compared with the substantial literature on cross-sectional zero-inflated count data, few papers have discussed the modeling of clustered, correlated observations, such as occurs with longitudinal data. Dobbie and Welsh (2001) applied marginal models using the generalized estimating equations approach for both parts of a hurdle model. If a within-subject effect is the focus of the study, a random effect approach is natural. Hall (2000) extended the Lambert (1992) zero-inflated Poisson (ZIP) model to handle longitudinal data, adding a random effect to account for the within-subject dependence in the Poisson state. However, Hall's model does not have a random effect for the part of the model determining the zero inflation. In contrast, Yau and Lee (2001) proposed adding a pair of uncorrelated normal random effects for the two components of a hurdle model. They used a penalized quasi-likelihood (PQL) approach for model fitting.

When the response is observed at several occasions, a high positive outcome at one time may increase the probability of a positive outcome at another time. These two processes are likely correlated and may be influenced by covariates in similar or in different ways. It makes sense to allow correlated random effects in a model, which then requires a more complex fitting process. In addition, the nonzero response may be overdispersed with respect to a truncated Poisson distribution and a truncated negative binomial distribution may be more appropriate. In this article, we develop correlated random effects models. In model fitting, Breslow and Lin (1995) showed that PQL estimators can be biased and inconsistent for highly non-normal (e.g., binary) responses when the random effects have large variance. Rather than PQL, we use parametric and nonparametric maximum likelihood (NPML) for model fitting. For a Poisson hurdle model, when the two parts of the hurdle model have the same covariates, we consider a special type of model that can be used to test for zero inflation. In addition, we consider a simpler approach using a single model – a cumulative logit model with random effects.

Section 2 briefly reviews the hurdle model and zero-inflated Poisson model and discusses advantages of the hurdle model. Section 3 introduces a hurdle model with random effects and discusses model fitting. The random effect cumulative logit model is introduced for this context in Section 4. Section 5 uses two examples to illustrate the methods.

2 Zero-inflated count data models

This section gives a brief overview of hurdle models and ZIP models for cross-sectional data.

2.1 Hurdle models

The hurdle model, proposed by Mullahy (1986), uses a two-stage modeling process. The first stage models the binary variable that measures whether the response falls below or above the hurdle. The second stage uses a truncated model to explain the observations above the hurdle. In the zero-inflated count data problem, the hurdle is zero. For response variable Y, let y_i denote the observation for subject $i, i = 1, \dots, n$. Suppose that the first part of the process is governed by a probability mass function g_1 and that $\{Y_i | Y_i > 0\}$ follows a truncated-at-zero probability mass function g_2, such as a truncated Poisson or negative binomial. The complete distribution is

$$P(Y_i = 0) = g_1(0) \tag{2.1}$$

$$P(Y_i = j) = (1 - g_1(0)) \frac{g_2(j)}{1 - g_2(0)} \quad j = 1, 2, \dots \tag{2.2}$$

This generalizes to models with explanatory variables, in which those affecting the first stage may not be the same as those affecting the second stage. Let $P(Y_i > 0) = p_i$ and $P(Y_i = 0) = 1 - p_i$. We use a logistic regression model for p_i and a log linear model for the mean μ_i of the untruncated g_2 distribution,

$$\text{logit}(p_i) = x'_{1i}\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\mu_i) = x'_{2i}\boldsymbol{\beta}_2$$

The likelihood function is

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^{n} (1 - p_i(\boldsymbol{\beta}_1))^{I(y_i=0)} \left[p_i(\boldsymbol{\beta}_1) \frac{g_2(y_i; \mu_i(\boldsymbol{\beta}_2))}{1 - g_2(0; \mu_i(\boldsymbol{\beta}_2))} \right]^{1-I(y_i=0)} \tag{2.3}$$

where $I(\cdot)$ is the indicator function. If $(1 - p_i) > g_2(0)$ for every i, the model represents zero inflation. If $(1 - p_i) < g_2(0)$ for every i, the model represents zero deflation.

The log likelihood factors into two terms,

$$\ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \ell_1(\boldsymbol{\beta}_1) + \ell_2(\boldsymbol{\beta}_2)$$

where,

$$\begin{aligned}\ell_1(\boldsymbol{\beta}_1) &= \sum_{y_i=0}[\log(1-p_i(\boldsymbol{\beta}_1))] + \sum_{y_i>0}[\log p_i(\boldsymbol{\beta}_1)] \\ &= \sum_{y_i>0} \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 - \sum_{i=1}^{n}\log(1+\mathrm{e}^{\boldsymbol{x}'_{1i}\boldsymbol{\beta}_1})\end{aligned}$$

is the log likelihood function for the binary process and

$$\ell_2(\boldsymbol{\beta}_2) = \sum_{y_i>0}[\log g_2(y_i;\ \mu(\boldsymbol{\beta}_2)) - \log(1-g_2(0;\ \mu(\boldsymbol{\beta}_2)))]$$

is the log likelihood function for the truncated model. One can obtain maximum likelihood (ML) estimates by separately maximizing the two terms.

2.2 Zero-inflated count models

An alternative approach for modeling zero-inflated data is the zero-inflated count model proposed by Lambert (1992). This model assumes that data are from a mixture of a regular count distribution, such as the Poisson distribution, and a degenerate distribution at zero. For a ZIP model, it is assumed that for subject i,

$$Y_i \sim \begin{cases} 0, & \text{with probability } 1-\phi_i \\ \text{Poisson } (\lambda_i), & \text{with probability } \phi_i \end{cases}$$

The probability distribution has

$$P(Y_i = 0) = (1-\phi_i) + \phi_i\,\mathrm{e}^{-\lambda_i} \tag{2.4}$$

$$P(Y_i = j) = \phi_i \frac{\mathrm{e}^{-\lambda_i}\lambda_i^j}{j!} \quad j = 1, 2, \ldots \tag{2.5}$$

With explanatory variables, the parameters are modeled by

$$\text{logit}\,(\phi_i) = \boldsymbol{x}'_{1i}\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\lambda_i) = \boldsymbol{x}'_{2i}\boldsymbol{\beta}_2$$

The EM algorithm or the Newton–Raphson method can be used to obtain the ML estimates. Compared with the hurdle model, this model is more complex to fit, as the model components must be fitted simultaneously.

2.3 Hurdle model versus ZIP model

The ZIP model is suitable only for handling zero inflation. However, the hurdle model is also suitable for modeling zero deflation. In fact, when a data set is zero deflated at a level of a factor, the estimate of the corresponding parameter in the first part of the ZIP model is ∞, so that the fit has no zero inflation at that level. The hurdle model does not have this problem.

We used two simple simulations to study this potential problem with the ZIP model. The first simulation assumed a hurdle model with Poisson g_2, at which there was zero deflation at one setting of the predictor but the entire data set (ignoring the covariate) tended to be zero inflated. Not surprisingly, the estimate of the predictor for fitting the ZIP model was highly unstable (Min and Agresti, 2004). However, when we used both van den Broek (1995) and Jansakul and Hinde (2002) score tests on our 1000 simulated data sets, all of them revealed evidence of zero inflation (P-value < 0.05 in each case). This simulation study tells us that even when a test shows significant evidence of zero inflation, the ZIP model may still not be suitable to fit the data.

More relevantly, the second simulation assumed a ZIP model, for which both models are valid. This used a case simulated by Lambert (1992);

$$\begin{aligned} \text{logit}(\phi_i) &= 1.5 - 2x_i \\ \log(\lambda_i) &= 1.5 - 2x_i \end{aligned}$$

We generated 1000 data sets, in which each data set had $n = 200$ observations. The covariate x_i is binary, taking value 0 for 100 cases and 1 for the other 100 cases. With these choices, on average, $\sim$51% of the responses were zeros, and 22% of those zeros were generated by the Poisson distribution.

For this assumed ZIP model, the hurdle model with Poisson g_2 also holds. The second part of the hurdle model has the same parameters as the second part of the ZIP model. Denote the first part of the hurdle model by $\text{logit}(p_i) = \beta^*_{10} + \beta^*_{11}x_i$. When $x_i = 0$,

$$P(Y_i = 0) = \frac{1}{1 + \exp(1.5)} + \frac{\exp(1.5)}{1 + \exp(1.5)} \exp(-\mathrm{e}^{1.5})$$

This equals $[1 + \exp(\beta^*_{10})]^{-1}$ in the hurdle model, so $\beta^*_{10} = 1.44$. When $x_i = 1$,

$$P(Y_i = 0) = \frac{1}{1 + \exp(-0.5)} + \frac{\exp(-0.5)}{1 + \exp(-0.5)} \exp(-\mathrm{e}^{-0.5})$$

This equals $[1 + \exp(\beta^*_{10} + \beta^*_{11})]^{-1}$ in the hurdle model, so $\beta^*_{11} = -3.01$.

Table 1 shows the results for this second simulation. Although the working model was a ZIP model, parts of some of the simulated data sets were not zero inflated. As estimates are sometimes unstable with the ZIP model (e.g., when a sample had zero deflation at a setting of x_i), we report the medians of the estimates and the SE values. In fact, instability often occurred for estimating β_{11} (the predictor effect in the logit component of the model) with that model, so the table shows only medians for estimates of that parameter for that model. We see that the hurdle model performed about the same as the ZIP model for the other parameters. However, the hurdle model performed better for estimating β_{11}.

Table 1 Comparing the estimated parameters of the ZIP model and the hurdle model for the simulated data sets from a hurdle working model

	β_{10}	β_{11}	β_{20}	β_{21}
ZIP Model				
Parameters	1.50	−2.0	1.5	−2.0
Mean estimate	1.461	–	1.500	−2.114
Median estimate	1.442	−1.865	1.504	−2.114
SE estimate[a]	0.266	–	0.057	0.460
Mean SE[b]	0.267	–	0.054	0.431
Median SE[b]	0.262	0.682	0.054	0.396
Hurdle Model				
Parameters	1.44	−3.01	1.5	−2.0
Mean estimate	1.399	−3.003	1.500	−2.106
Median estimate	1.386	−2.983	1.503	−2.033
SE estimate[a]	0.252	0.370	0.057	0.461
Mean SE[b]	0.253	0.371	0.054	0.452
Median SE[b]	0.250	0.369	0.054	0.427

[a] The SE estimate is the standard deviation of the 1000 estimates of the parameter.
[b] Mean (median) SE is the average of (or the median of) the 1000 estimated standard errors.

The simulation studies tell us that the ZIP model may be unreliable in fitting zero-inflated count data, even for simple cross-sectional data. Therefore, in the remainder of the paper, we mainly discuss the extension of the hurdle model to repeated measures.

3 Hurdle models with random effects

3.1 Model specification

Now we extend the hurdle model to clustered, correlated counts. Let y_{ij} be observation $j(j = 1, \dots, t_i)$ for subject (or cluster) i $(i = 1, \dots, n)$. Define

$$u_{ij} = \begin{cases} 0, & \text{if } y_{ij} = 0 \\ 1, & \text{if } y_{ij} > 0 \end{cases}$$

and let $p_{ij} = P(y_{ij} > 0)$. Suppose the positive count response follows a truncated count distribution with probability mass function g having mean μ_{ij} for the untruncated count distribution. Let $\boldsymbol{b}_i = (\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i})'$ be random effects designed to account for within-subject correlation. Conditional on $\boldsymbol{b}_i$, we assume that

$$\text{logit}(p_{ij}) = \boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + \boldsymbol{z}'_{1ij}\boldsymbol{b}_{1i} \tag{3.1}$$

$$\log(\mu_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + \boldsymbol{z}'_{2ij}\boldsymbol{\beta}_{2i} \tag{3.2}$$

where $\boldsymbol{x}_{kij}$ and $\boldsymbol{z}_{kij}$ are covariate vectors pertaining to the fixed effects $\boldsymbol{\beta}_k$ and the random effects $\boldsymbol{b}_{ki}$. In practice, the simple random intercept form of models is often adequate, in which $\boldsymbol{b}_{1i} = b_{1i}$ and $\boldsymbol{b}_{2i} = b_{2i}$ are univariate and $z_{1ij} = z_{2ij} = 1$.

When the response is observed at repeated times, as in longitudinal studies, the response at one time may be positively correlated with the response at another time. One can tie the two parts of the model together by assuming that the random effects are jointly normal and possibly correlated,

$$\boldsymbol{b}_i = \begin{pmatrix} b_{i1} \\ b_{i2} \end{pmatrix} \sim \mathbf{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{bmatrix} \right)$$

where Σ, Σ_{11} and Σ_{22} are unknown positive-definite matrices. Let $\boldsymbol{\psi}$ represent the unknown parameters, $\boldsymbol{\psi} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \Sigma)$. The marginal log likelihood for the hurdle random effect model is:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\psi})$$

where

$$\begin{aligned} L_i(\boldsymbol{\psi}) &= \int \left[\prod_{j=1}^{t_i} (1 - p_{ij})^{1-u_{ij}} \left(p_{ij} \frac{g(y_{ij})}{1 - g(0)} \right)^{u_{ij}} \right] \phi(\boldsymbol{b}_i) \mathrm{d}\boldsymbol{b}_i \\ &= \int \left[\prod_{j=1}^{t_i} f_1(u_{ij} \mid \boldsymbol{b}_{1i}) f_2(y_{ij}, u_{ij} \mid \boldsymbol{b}_{2i}) \right] \phi(\boldsymbol{b}_i) \mathrm{d}\boldsymbol{b}_i \end{aligned}$$

and ϕ denotes the normal density function for the random effects.

3.2 A special case of the hurdle model

In the univariate response case, Heilbron (1994) defined a type of hurdle model as a compatible two-part model. This type of model requires that g_1 and g_2 in Section 2.1 have identical distribution forms and overdispersion parameters. It also assumes that the covariates and the link functions for modeling the means of the distributions are the same for the two parts, that is, $\eta(\mu_1) = \boldsymbol{x}'\boldsymbol{\beta}_1$ and $\eta(\mu_2) = \boldsymbol{x}'\boldsymbol{\beta}_2$. If $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, the model would reduce to a standard generalized linear model. For covariate x_k, the difference $\beta_{2k} - \beta_{1k}$ can be used to explain its effect on the zero inflation. For the Poisson distribution, a log linear model for the mean of the untruncated count distribution is equivalent to using a complementary log–log link for the first part of the model, which is

$$\log(-\log(1 - p_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}_1$$

By assuming that μ_1 is a function of μ_2, such as $\mu_1 = \gamma_1 \mu_2^{\gamma_2}$ $(\gamma_1 > 0, \gamma_2 \geq 0)$, Heilbron proposed a zero-altered model that has the form $\boldsymbol{\beta}_1 = \log(\gamma_1) + \gamma_2 \boldsymbol{\beta}_2$. Through comparing to the simpler model with $\gamma_1 = 1$ and $\gamma_2 = 1$, one can test whether a standard Poisson model is sufficient to fit the data. For $\gamma_2 = 1$, if $\gamma_1 < 1$, the data are zero inflated, as it is equivalent to $\mu_1 < \mu_2$, whereas if $\gamma_1 > 1$, the data are zero deflated.

The zero-altered model extends to the repeated measures setting. For simplicity, we let $\boldsymbol{b}_i \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$ be a subject-specific random effect for both parts of the hurdle model. Conditional on $\boldsymbol{b}_i$, we assume

$$\log(-\log(1 - p_{ij})) = \gamma_1' + \gamma_2(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) + \boldsymbol{b}_i \tag{3.3}$$

$$\log(\mu_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{b}_i \tag{3.4}$$

We call this the zero-altered Poisson random effect model. Most papers dealing with clustered, correlated zero-inflated count data test the existence of zero inflation for the data at different occasions separately. Setting $\gamma_2 = 1$, through testing whether $\gamma_1' = 0$, one can test the existence of zero inflation for clustered count data. If $\gamma_1' < 0$, the data are zero inflated; if $\gamma_1' > 0$, the data are zero deflated. One can use a likelihood ratio test to conduct these tests. When we set $\gamma_2 = 1$, this model also has the convenient property of a single set of effects $\boldsymbol{\beta}$. For instance, to compare different groups that are levels of the explanatory variables, one can use $\hat{\boldsymbol{\beta}}$ directly, whereas for the general hurdle model with random effects one needs to average results from the two components of the model to make an unconditional comparison [e.g., to estimate E(Y) for the groups].

3.3 ML model fitting with normal random effects

To fit a hurdle model with random effects, one first obtains the marginal likelihood by integrating out the random effects. These integrals are analytically intractable, so numerical or stochastic approximation of them is needed. There are many methods to approximate the ML estimate for generalized linear mixed models (GLMM) (Fahrmeir and Tutz, 2001; McCulloch and Searle, 2001) such as Gauss–Hermite quadrature the Monte Carlo EM algorithm, Markov chain Monte Carlo, PQL and Laplace approximations. The first three methods have the advantage that they converge to the ML estimate as they are applied more finely. As the simple random intercept form of models is often adequate in practice, we only discuss the case with $\boldsymbol{b}_{1i} = b_{1i}$ and $\boldsymbol{b}_{2i} = b_{2i}$ univariate and $z_{1ij} = z_{2ij} = 1$. With univariate random intercepts, numerical approximation using Gauss–Hermite quadrature, which approximates the integral by a finite sum is adequate.

Let

$$f(\boldsymbol{b}_i) = \prod_{j=1}^{t_i} [f_1(u_{ij} \mid \boldsymbol{b}_{1i}) f_2(y_{ij}, u_{ij} \mid \boldsymbol{b}_{2i})]$$

We assume that

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

and let $\boldsymbol{L}$ be the lower triangular Cholesky factor of Σ. We transform $\boldsymbol{b}_i = \sqrt{2}\boldsymbol{L}\boldsymbol{c}_{\mathrm{i}}$, where $\Sigma = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$. Then, the likelihood function for the ith

$$L_i(\boldsymbol{\psi}) = \int f(\boldsymbol{b}_i)\frac{1}{|2\pi\Sigma|^{1/2}}\exp\left(-\frac{1}{2}\boldsymbol{b}_i^{\mathrm{T}}\Sigma^{-1}\boldsymbol{b}_i\right)\mathrm{d}\boldsymbol{b}_i$$
$$= \frac{1}{\pi}\int f(\sqrt{2}\boldsymbol{L}\boldsymbol{c}_i)\exp(-\boldsymbol{c}_i^{\mathrm{T}}\boldsymbol{c}_i)\mathrm{d}\boldsymbol{c}_i$$

The Gauss–Hermite approximation using m quadrature points for each dimension is

$$L_i^{\mathrm{GH}}(\boldsymbol{\psi}) = \sum_{l_1=1}^{m} v_{l_1}^{(1)} \sum_{l_2=1}^{m} v_{l_2}^{(2)} f(c_{l_1}^{(1)}, c_{l_2}^{(2)})$$

where $v_{l_k}^{(k)} = \pi^{-1/2}w_{l_k}^{(k)}$, and $c_{l_k}^{(k)}$ and $w_{l_k}^{(k)}$ are the node k and weight k $(k = 1, 2)$ of the univariate Gauss–Hermite integration of order m.

To maximize this approximation for the likelihood function, we use an approximate version of the Fisher scoring method (Green, 1984; Raudenbush *et al.*, 2000) to obtain $\hat{\boldsymbol{\psi}}$. Let $S(\boldsymbol{\psi})$ score vector, which is approximated as

$$S(\boldsymbol{\psi}) \approx \sum_{i=1}^{n} S_i^{\mathrm{GH}}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \frac{\partial \log L_i^{\mathrm{GH}}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$$
$$= \sum_{i=1}^{n} \frac{1}{L_i^{\mathrm{GH}}(\boldsymbol{\psi})} \sum_{l_1=1}^{m} v_{l_1}^{(1)} \sum_{l_2=1}^{m} v_{l_2}^{(2)} f(y_i, \boldsymbol{c}^{(l)};\ \boldsymbol{\psi}) \frac{\partial \log f(y_i, \boldsymbol{c}^{(l)}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$$

where $\boldsymbol{c}^{(l)} = (c_{l_1}^{(1)}, c_{l_2}^{(2)})'$. The Fisher scoring method obtains the ML estimates by iteratively solving the equation $\boldsymbol{\psi}^{(t+1)} = \boldsymbol{\psi}^{(t)} + \boldsymbol{I}^{-1}(\boldsymbol{\psi}^{(t)})S(\boldsymbol{\psi}^{(t)})$ until $\boldsymbol{\psi}^{(t)}$ converges, where $\boldsymbol{I} = -\mathrm{E}[\sum_{i=1}^{n} \partial^2 \log L_i/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}']$. The second derivatives are usually difficult to calculate. Thus, we used an approximate scoring procedure with $\boldsymbol{I} \approx \sum_{i=1}^{n} S_i(\boldsymbol{\psi})S_i(\boldsymbol{\psi})^{\mathrm{T}}$.

With univariate random effects, one can use the SAS procedure NLMIXED to fit this type of model as well as the random effect zero-inflated count model. SAS NLMIXED uses the adaptive Gauss–Hermite quadrature (Liu and Pierce, 1994; Pinheiro and Bates, 1995) to approximate the integrals, and the default maximization approach is the quasi-Newton method.

3.4 ML model fitting with a nonparametric approach

Section 3.3 assumed that $\phi(\boldsymbol{b}_i)$ is a bivariate normal probability density function. As severe misspecification of the random effect distribution could potentially bias parameter estimation, Aitkin (1999) suggested using an unspecified discrete distribution for the random effects. We extend his NPML method in this section for a bivariate random effect in the hurdle model.

We assume that ϕ is an unknown discrete distribution with K mass points $\boldsymbol{m} = (\boldsymbol{m}_1', \ldots, \boldsymbol{m}_K')'$ and corresponding probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, where $\boldsymbol{m}_k = (m_{1k}, m_{2k})'$, $k = 1, \ldots, K$. The log likelihood function is

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \left[\prod_{j=1}^{t_i} f(y_{ij};\ \boldsymbol{\beta} \mid \boldsymbol{m}_k) \right]$$

where

$$f(y_{ij};\ \boldsymbol{\beta} \mid \boldsymbol{m}_k) = f_1(u_{ij};\ \boldsymbol{\beta}_1 \mid m_{1k}) f_2(y_{ij}, u_{ij};\ \boldsymbol{\beta}_2 \mid m_{2k})$$

This type of finite mixture model can be related to a latent class model (Aitkin and Rubin, 1985), which is useful for model fitting. Suppose that d_{ik} is an indicator that represents whether $\boldsymbol{y}_i$ is drawn from the kth latent group, $\sum_k d_{ik} = 1$. Assume that $(\boldsymbol{y}_i \mid d_i, \boldsymbol{\beta}, \boldsymbol{m}_k)$ are independently distributed with densities

$$\sum_{k=1}^{K} d_{ik} f(\boldsymbol{y}_i;\ \boldsymbol{\beta} | \boldsymbol{m}_k) = \prod_{k=1}^{K} f(\boldsymbol{y}_i;\ \boldsymbol{\beta} \mid \boldsymbol{m}_k)^{d_{ik}}$$

where $f(\boldsymbol{y}_i;\ \boldsymbol{\beta} \mid \boldsymbol{m}_k) = \prod_{j=1}^{t_i} f(y_{ij};\ \boldsymbol{\beta} \mid \boldsymbol{m}_k)$. Assume that $(d_{ik} \mid \boldsymbol{\pi})$ are i.i.d. with multinomial distribution $\prod_{k=1}^{K} \pi_k^{d_{ik}}$. Treating $\{d_{ik}\}$ as missing data, the EM algorithm can be used to estimate this finite mixture model. The complete log likelihood function is

$$\ell_c(\boldsymbol{\psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} d_{ik} \left[\sum_{j=1}^{t_i} \log f(y_{ij};\ \boldsymbol{\beta} \mid \boldsymbol{m}_k) + \log \pi_k \right]$$

In iteration t, the E-step calculates the expectation of the complete log likelihood, that is,

$$\begin{aligned} \mathrm{E}[\ell_c(\boldsymbol{\psi} \mid \boldsymbol{\psi}^{(t)})] &= \sum_{i=1}^{n} \sum_{j=1}^{t_i} \sum_{k=1}^{K} w_{ik}^{(t)} \log f(y_{ij};\ \boldsymbol{\beta} \mid \boldsymbol{m}_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(t)} \log \pi_k \\ &= h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1) + h_2(\boldsymbol{\beta}_2, \boldsymbol{m}_2) + h_3(\boldsymbol{\pi}) \end{aligned}$$

where

$$h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1) = \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{k=1}^{K} w_{ik}^{(t)} \log f_1(u_{ij}; \boldsymbol{\beta}_1 | m_{1k})$$
$$h_2(\boldsymbol{\beta}_2, \boldsymbol{m}_2) = \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{k=1}^{K} w_{ik}^{(t)} \log f_2(y_{ij}, u_{ij}; \boldsymbol{\beta}_2 | m_{2k})$$
$$h_3(\boldsymbol{\pi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \log \pi_k$$

and

$$w_{ik}^{(t)} = \frac{\pi_k^{(t)}\left[\prod_{j=1}^{t_i} f(y_{ij}; \boldsymbol{\beta}^{(t)} | \boldsymbol{m}_k^{(t)})\right]}{\sum_{l=1}^{K} \pi_l^{(t)}\left[\prod_{j=1}^{t_i} f(y_{ij}; \boldsymbol{\beta}^{(t)} | \boldsymbol{m}_l^{(t)})\right]}$$

being the posterior mean of d_{ik}. In the M-step, we maximize $\mathrm{E}[\ell_c(\boldsymbol{\psi}|\boldsymbol{\psi}^{(t)})]$ with respect to $\boldsymbol{\psi}$ to obtain $\boldsymbol{\psi}^{(t+1)}$. As $(\boldsymbol{\beta}_1, \boldsymbol{m}_1)$, $(\boldsymbol{\beta}_2, \boldsymbol{m}_2)$ $\boldsymbol{\pi}$ are in three separate terms, we can maximize $h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1)$, $h_2(\boldsymbol{\beta}_2, \boldsymbol{m}_2)$ and $h_3(\boldsymbol{\pi})$, separately. When maximizing with respect to $\boldsymbol{\pi}$, we need to take the constraint $\sum_{k=1}^{K} \pi_k = 1$ into consideration. Solving the equations

$$\frac{\partial}{\partial \pi_k}\left[h_3(\boldsymbol{\pi}) - \lambda\left(\sum_{l=1}^{K} \pi_k - 1\right)\right] = \frac{1}{\pi_k} - \lambda = 0$$

yields

$$\pi_k^{(t+1)} = \sum_{i=1}^{n} w_{ik}^{(t)}/n$$

The maximization with respect to $(\boldsymbol{\beta}_1, \boldsymbol{m}_1)$ is a weighted version of binomial distribution ML estimation with logit link. Let $\mathrm{logit}(p_{ijk}) = \boldsymbol{x}_{1ij}'\boldsymbol{\beta}_1 + \boldsymbol{m}_{1k}$. We can get $\boldsymbol{\beta}_1^{(t+1)}$ and $\boldsymbol{m}_1^{(t+1)}$ by solving

$$\frac{\partial h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1)}{\partial \boldsymbol{\beta}_1} = \sum_{i=1}^{n}\sum_{j=1}^{t_i}\sum_{k=1}^{K} w_{ik}^{(t)}(u_{ij} - p_{ijk})\boldsymbol{x}_{1ij} = 0$$
$$\frac{\partial h_1(\boldsymbol{\beta}_1, \boldsymbol{m}_1)}{\partial m_{1k}} = \sum_{i=1}^{n}\sum_{j=1}^{t_i} w_{ik}^{(t)}(u_{ij} - p_{ijk}) = 0 \quad k = 1, \ldots, K$$

The maximization with respect to $(\boldsymbol{\beta}_2, \boldsymbol{m}_2)$ is a weighted version of ML estimation of a truncated Poisson model or a truncated negative binomial model. Convergence of the

EM algorithm can be determined by the Euclidean norm of the difference in parameter estimates. In order to avoid a local maximum, trying different starting values is recommended. Standard errors of the fixed effects can be obtained by calculating the inverse of the observed information matrix (Louis, 1982).

For a given choice of the number K of mass points, the estimated maximized log-likelihood is

$$\ell_K(\hat{\boldsymbol{\psi}}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \hat{\boldsymbol{\pi}}_k \left[\prod_{j=1}^{t_i} f(y_{ij}; \hat{\boldsymbol{\beta}} | \hat{\boldsymbol{m}}_k) \right]$$

We define the deviance difference, comparing this model to the simpler nonmixture model, by

$$\mathrm{dev}_K = 2[\ell_K(\hat{\boldsymbol{\psi}}) - \ell_1(\hat{\boldsymbol{\psi}})]$$

where $\ell_1(\hat{\boldsymbol{\psi}})$ is the estimated maximum log likelihood function for $\sum_{i=1}^{n} t_i$ independent responses. Although this does not give a formal significance test (since the simpler model is on the boundary of the parameter space), the support size of K can be estimated by starting with $K = 2$ and increasing K until the change in the deviance is small.

4 Cumulative logit models with random effects

In a methadone programme evaluation study (Saei *et al.*, 1996), one of the response variables was the number of crimes committed during the previous three months. Saei *et al.* (1996) suggested grouping the possible count outcomes into K ordered categories and applying an ordinal response model with random effects. They proposed a cumulative probit model with random effects and used the PQL approach to estimate the parameters. We propose a cumulative logit model with random effects and use parametric ML for model fitting.

Let $Y_{ij,g}$ be the grouped response variable for observation j on subject i. The threshold model for an ordinal response posits an unobservable variable Z, such that one observes $Y_{ij,g} = k$ category if Z is between θ_{k-1} and θ_k. Suppose that Z has a cumulative Q1
distribution function $G(z - \eta)$, where η is related to explanatory variables by

$$\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i$$

for a vector $\boldsymbol{b}_i \sim N(0, \Sigma)$ of random effects that account for within-subject correlation. Then,

$$P(Y_{ij,g} \le k) = P(Z \le \theta_k) = G(\theta_k - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - \boldsymbol{z}'_{ij}\boldsymbol{b}_i)$$

The inverse of the CDF of G serves as the link function. Assuming that G is logistic leads to a logit model for the cumulative probabilities with random effects.

In applications with zero-inflated count data, one would take the first category to be the zero outcome, and then treat each other outcome as a separate category, or group count values together to form the other $K-1$ categories. When grouping the count values together to form the K categories, our simulation studies suggest that when the number of groups is too small, one will lose some efficiency. We suggest that the grouping size should be at least four. However, using more than four or five categories does not increase efficiency much, and it has the disadvantage that one needs to estimate more parameters.

The model has the form

$$\text{logit}[P(Y_{ij,g} \leq k;\ \boldsymbol{x})] = \eta_{ijk} = \theta_k - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - \boldsymbol{z}'_{ij}\boldsymbol{\beta}_i, \quad k = 1, 2, \ldots, K-1 \tag{4.1}$$

The probability that $Y_{ij,g}$ takes value k is

$$\pi_{ij,k} = P(Y_{ij,g} = k) = \frac{1}{1+\exp(-\eta_{ijk})} - \frac{1}{1+\exp(-\eta_{ij,k-1})} \quad k = 1, 2, \ldots, K$$

where $\eta_{ij0} = -\infty$. For subject i at occasion j, define $y_{ijk} = 1$ if $Y_{ij,g} = k(k = 1, 2, \ldots, K)$ Q2
and $y_{ijk} = 0$ otherwise. Then $\boldsymbol{y}_{ij} = (\boldsymbol{y}_{ij1}, \ldots, \boldsymbol{y}_{ijK})'$ is a K-dimensional vector following a multinomial $\prod_{k=1}^{K} \pi_{ij,k}^{y_{ijk}}$ distribution. Let $f(\boldsymbol{y}_{ij};\ \boldsymbol{\beta}|\boldsymbol{b}_i)$ be the multinomial probability mass function and ϕ be the multivariate normal density function with mean $\mathbf{0}$ and covariance $\sum$. The marginal log likelihood for the cumulative logit random effect model is:

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n} \int \left[\prod_{j=1}^{t_i} f(\boldsymbol{y}_{ij};\ \boldsymbol{\beta}|\boldsymbol{b}_i)\right] \phi(\boldsymbol{b}_i)\,\mathrm{d}\boldsymbol{b}_i$$

This is similar to the log likelihood function in Section 3. One can use the SAS procedure NLMIXED to fit this model with ML. Hartzel *et al.* (2001) provided a nonparametric approach for the random effects in ML model fitting.

This model has the simplicity of a single equation to handle the clump at zero and the positive outcomes. Elements of $\boldsymbol{\beta}$ describe effects overall, rather than conditional on the response being positive. This is an important advantage. For instance, to compare different groups that are levels of the explanatory variables, one can use $\hat{\boldsymbol{\beta}}$ directly, whereas for general two-part hurdle models one needs to average results from the two components of the model to make an unconditional comparison.

5 Applications

5.1 A pharmaceutical study

Our main example refers to a data set shown to us by a pharmaceutical company. Unfortunately, we are unable to use their original data or discuss details of the study because of the company's confidentiality restrictions. We changed some numbers in the data set, while keeping the basic structure of the data, such as its zero inflation. (The data set and related SAS code are available at the journal's web site.)

One aspect of this study dealt with comparing two treatments for a particular disease in terms of the number of episodes of a certain side effect. The study had 118 patients, with 59 randomly allocated to receive treatment A (TRT1) and the other 59 receiving treatment B (TRT2). The number of side effect episodes was measured at each of six visits. Of the observations $\sim 83\%$ were zeros. Table 2 shows the frequencies of the side effect for treatments A and B. As the count data vary with exposure time between visits, we incorporated time-between-visit (defined as Time) as a covariate in the model.

First, we fitted an ordinary Poisson GLMM and a type II negative binomial GLMM. The density function of y_i for the negative binomial model is

$$g(y_i;\ \mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}$$

with dispersion parameter $\alpha > 0$. As $\alpha \to 0$, the negative binomial distribution converges to the Poisson distribution. For both the Poisson GLMM and the negative binomial GLMM, the model is

$$\log(\mu_{ij}) = \beta_0 + \beta_1\ \mathrm{TRT2} + \beta_2\ \log(\mathrm{Time}) + b_i$$

where b_i is assumed to have a $N(0, \sigma^2)$ distribution. Table 3 shows estimates for these models. The estimated $\hat{\alpha}$ in the negative binomial random effect model or the reduction in log likelihood when compared with the Poisson GLMM suggest that the Poisson GLMM is inadequate. For a formal test of the hypothesis that $\alpha = 0$, the likelihood ratio statistic comparing the models has an asymptotic null distribution, that is, 1/2 : 1/2 mixture of a $\chi_1{}^2$-distribution and a point-mass at 0 (Self and Liang, 1987). The test statistic equals 19, giving strong evidence of zero inflation.

We then fitted a zero-altered random effect Poisson model, which has the form

$$\begin{aligned} \log(1 - \log(1 - p_{ij})) &= \gamma_1 + \beta_0 + \beta_1\ \mathrm{TRT2} + \beta_2\ \log(\mathrm{Time}) + b_i \\ \log(\mu_{ij}) &= \beta_0 + \beta_1\ \mathrm{TRT2} + \beta_2 \log(\mathrm{Time}) + b_i \end{aligned}$$

where b_i again has a $N(0, \sigma^2)$ distribution. The likelihood ratio test of $\mathrm{H}_0 : \gamma_1 = 0$ has test statistic = 21.2 with df = 1, also giving strong evidence of zero inflation.

Both the negative binomial random effect model and the zero-altered random effect model showed that the ordinary Poisson GLMM is inadequate. The estimated

Table 2 Side effect frequencies in treatment A and treatment B

	Frequencies						
Treatment	0	1	2	3	4	5	6
A	312	30	11	0	1	0	0
B	278	39	20	6	7	2	2
Total	590	69	31	6	8	2	2

Table 3 Parameter estimates for the Poisson negative binomial and zero-altered random effect models for modeling side effects

Parameter	Poisson		Negative binomial		Zero-altered	
	Estimate	SE	Estimate	SE	Estimate	SE
β_0	−3.492	0.454	−3.350	0.524	−2.557	0.472
β_1(TRT2)	0.947	0.373	0.941	0.372	0.833	0.319
β_2(log(Time))	0.238	0.114	0.212	0.147	0.220	0.113
Σ	1.535	0.204	1.488	0.211	1.244	0.194
α	–	–	0.766	0.275	–	–
γ_1	–	–	–	–	−0.771	0.169
$-2\ell(\hat{\psi})$	852.1	–	833.1	–	830.9	–

parameter comparing the treatments is $\hat{\boldsymbol{\beta}}_1 = 0.833$ in the zero-altered random effect model and $\hat{\boldsymbol{\beta}}_1 = 0.941$ in the negative binomial random effect model, each being about 2.5 standard errors. These suggest that treatment B has a higher probability of the side effect and a higher number of episodes than treatment A.

The Poisson random effect hurdle model has the form

$$\text{logit}(p_{ij}) = \beta_{10} + \beta_{11}\text{TRT2} + \beta_{12}\log(\text{Time}) + b_{1i}$$

$$\log(\mu_{ij}) = \beta_{20} + \beta_{21}\text{TRT2} + \beta_{22}\log(\text{Time}) + b_{2i}$$

where (b_{1i}, b_{2i}) have a bivariate normal distribution. Table 4 shows the ML estimates. Compared to the zero-altered model and the negative binomial model, this gives us the extra information that the time-between-visits seems to have little effect on the probability of getting the side effect, whereas it has a substantial effect on the number of episodes. We also fitted a negative binomial random effect hurdle model. It has very similar maximized log likelihood value as the Poisson hurdle model and $\hat{\alpha}$ close to zero. Therefore, we do not show its estimates here.

We also used the nonparametric approach to fit the Poisson random effect hurdle model. We obtained similar results as the with the ordinary ML approach. In the

Table 4 Parameter estimation of the random effect hurdle models for the numbers of side effects

Parameter	ML		NPML	
	Estimate	SE	Estimate	SE
β_{10}(Intercept)	−2.874	0.622	−2.813	0.576
β_{11}(TRT2)	0.895	0.417	0.958	0.335
β_{12}(log(Time))	0.021	0.186	0.022	0.185
Σ_1	1.647	0.251	–	–
β_{20}(Intercept)	−2.844	0.735	−2.880	0.619
β_{21}(TRT2)	0.963	0.352	0.898	0.294
β_{22}(log(Time))	0.540	0.192	0.494	0.188
Σ_2	0.706	0.248	–	–
ρ	0.848	0.204	–	–
$-2\ell(\hat{\psi})$	818.3	–	809.3	–

NPML analysis of the Poisson hurdle model allowing correlated random effects, $-2\ell_1(\hat{\psi}) = 888.9$ with $K = 1$ support point; $-2\ell_2(\hat{\psi}) = 816.2$ with $K = 2$ points; $-2\ell_3(\hat{\psi}) = 809.3$ with $K = 3$ points and $-2\ell_4(\hat{\psi}) = 809.1$ with $K = 4$ points. We used $K = 3$, for which the estimated mass points are $\hat{\boldsymbol{m}}_1 = (0.34, -0.99)'$, $\hat{\boldsymbol{m}}_2 = (2.18, 1.20)'$ and $\hat{\boldsymbol{m}}_3 = (-2.15, 1.69)'$ with $\hat{\boldsymbol{\pi}} = (0.34, 0.24, 0.42)'$. Table 4 shows the estimates, which are substantively similar to ordinary ML. Again, they show that time-between-visits is more important for determining the number of episodes than whether they occur.

This example had only seven possible outcomes (Table 2). To use the cumulative logit model approach, we grouped the response variable into the five categories (0, 1, 2, 3, 4, >4). The cumulative logit model is

$$\text{logit}[P(Y_{ij,g} \leq k)] = \theta_k - \beta_1 \text{TRT2} - \beta_2 \log(\text{Time}) - b_i \quad k = 0, \ldots, 4$$

where $b_i \sim N(0, \sigma^2)$ accounts for within-subject correlation. Table 5 shows the ML estimates. Time-between-visits does not have a significant effect on the number of episodes. Like the zero-altered model, the model has the advantage of a single effect parameter for each predictor.

The hurdle random effect model suggested that treatment B has a higher probability of the side effect and a higher expected number of episodes than treatment A. The cumulative logit random effect model fitting confirms this conclusion, as $\hat{\boldsymbol{\beta}}_1 = 0.977$ has a standard error of 0.431. The estimated odds that the number of side effects falls below any fixed category with treatment A are exp(0.977) = 2.7 times the estimated odds for treatment B. This estimate has the same order of magnitude as the estimate from the binary part of the hurdle random effects model, which is not surprising since the probability modeled there is the first cumulative probability.

Table 6 summarizes − 2 log likelihood values for various fitted models. Some of the models are non-nested such as the Poisson hurdle random effect model and the cumulative logit random effect model. Therefore, we cannot simply compare their log likelihood values directly. For this example, the Poisson hurdle random effect model seems adequate. It has a relatively small − 2 log likelihood value, and it also provides the information that time-between-visits has little effect on the probability of getting the

Table 5 ML estimates for the cumulative logit random effect model for the numbers of side effects

Parameter	Estimate	SE
θ_0	3.311	0.629
θ_1	4.619	0.655
θ_2	5.901	0.703
θ_3	6.373	0.730
θ_4	7.554	0.846
β_1(TRT2)	0.977	0.43
β_2(log(Time))	0.153	0.181
σ	1.733	0.252
$-2\ell(\hat{\psi})$	817.2	

Table 6 Summary of $-2\ell(\hat{\psi})$ for ML fitting of various models

Model	$-2\ell(\hat{\psi})$	No. of parameters
Poisson GLMM	852.1	4
Negative binomial GLMM	833.1	5
Zero-altered random effect Poisson model	830.9	5
Poisson hurdle model with random effects (ML)	818.3	9
Poisson hurdle model with random effects (NPML $K=3$)	809.3	14
Cumulative logit model with random effects	817.2	8

side effect but considerable effect on the number of episodes. We cannot learn this by fitting the other models listed in this table.

5.2 An occupational injury prevention program study

We briefly mention a second example, from Yau and Lee (2001). They evaluated the effectiveness of an occupational injury prevention program used in the cleaning services of the studied Australian hospital. This pilot program used workplace risk assessment Teams (WRATS) intervention to attempt to reduce the expected number of manual handling injuries. The data set comprised injury counts from 137 cleaners who were present in pre- and post-WRATS intervention. Of the pre-WRATS observations ~52.6% were zero. Of the post-WRATS observations ~78.8% were zero. The explanatory variables include time, age, gender and the time of exposure variable.

Yau and Lee conducted overdispersion tests (Böhning *et al.*, 1997) separately on pre-WRATS and post-WRATS counts. They found overdispersion for the pre-WRATS data, and thus treated the data as zero-inflated count data. However, separate tests on each cross-sectional part of the data are less appealing than methods that recognize the repeated measures aspect of the data. A zero-altered random effect model does not find significant evidence of zero inflation (likelihood ratio test statistic $=1.7$ with df $=1$). An ordinary Poisson GLMM seems sufficient for these data. Unlike the ZIP model, the hurdle model does not require the data to be zero inflated. Yau and Lee fitted the two parts of a Poisson hurdle model with random effects separately and used the PQL approach for model fitting. We used ML to fit this model and found that the PQL approach does not approximate the ML estimates well. Through fitting the two components of the model jointly, we showed evidence of efficiency gains of fitting a correlated random effect model. A random effect cumulative logit model also fits well and gives simple summaries. Detailed analysis can be found in Min and Agresti (2004).

6 Summary

We have proposed a two-part random effect model, which has a binary component and a truncated Poisson or negative binomial component. When the two parts have the same covariates, a zero-altered random effect model can be used to test for zero inflation. The two-part model approach is also suitable for zero-deflated count data cases. When the response variable has relatively few count outcomes, a cumulative logit random effect model provides a simple way to handle this kind of problem. The ZIP random effect model requires that the data be zero inflated at every level of the covariates. This requirement is sometimes not realistic. Fitting a ZIP random effect model is also more complex than fitting a hurdle random effect model.

In general, for repeated measures of count data with zero inflation, the simpler models – the zero-altered random effect model or the cumulative logit random effect model – have the advantage of simplicity of interpretation. If one wants to estimate different covariate effects on zero responses and on nonzero responses, such as in checking, whether a predictor affects only the probability of a positive response, the hurdle model with random effects is more natural.

Acknowledgements

This work was partially supported by grants from NIH and NSF. The authors thank Dr Andy Lee for providing the data from the Yau and Lee (2001) paper.

References

Aitkin M (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117–28.

Aitkin M and Rubin DB (1985) Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society: Series B Methodological* **47**, 67–75.

Arulampalam W and Booth A (1997) Who gets over the training hurdle? a study of the training experiences of young men and women in Britain. *Journal of Population Economics* **10**, 197–217.

Böhning D, Dietz E and Schlattmann, P (1997) Zero-inflated count models and their applications in public health and social science. In, Rost J and Langeheine R, editors, *Applications of Latent Trait and Latent Class Models in the Social Sciences*. Münster: Waxmann. Q3

Breslow NE and Lin X (1995) Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Dobbie MJ and Welsh AH (2001) Modeling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics* **43**, 431–44.

Fahrmeir L and Tutz G (2001) *Multivariate statistical modelling based on generalized linear models*. 2nd edition. New York: Springer-Verlag.

Green PJ (1984) Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society: Series B Methodological* **46**, 149–92.

Hall D (2000) Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030–39.

Hartzel J, Agresti A and Caffo B (2001) Multinomial logit random effects models. *Statistical Modelling* **1**, 81–102.

Heilbron DC (1994) Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* **36**, 531–47.

Jansakul N and Hinde JP (2002) Score tests for zero-inflated Poisson models. *Computational Statistics and Data Analysis* **40**, 75–96.

Jiang J (1998) Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association* **93**, 720–29. Q4

Lambert D (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Liu Q and Pierce DA (1994) A note on Gauss–Hermite quadrature. *Biometrika* **81**, 624–29.

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B Methodological* **44**, 226–33.

Min Y and Agresti A (2004) Random effects models for repeated measures of zero-inflated count data. *Technical report no.* 2004-026, Department of Statistics, University of Florida.

McCulloch CE and Searle SR (2001) *Generalized, linear, and mixed models*, New York, Chichester: John Wiley & Sons.

Mullahy J (1986) Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–65.

Pinheiro JC and Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

Raudenbush SW, Yang M-L and Yosef M (2000) Maximum likelihood for generalized linear models with nested random effects via high-order Laplace approximation. *Journal of Computational and Graphical Statistics* **9**, 141–57.

Saei A, Ward J and McGilchrist CA (1996) Threshold models in a methadone programme evaluation. *Statistics in Medicine* **15**, 2253–60.

Self SG and Liang KY (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *Journal of the American Statistical Association* **82**, 605–10.

Shankar V, Milton J and Mannering F (1997) Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis and Prevention* **29**, 829–37.

van den Broek J (1995) A score test for zero inflation in a Poisson distribution. *Biometrics* **21**, 738–43.

Yau KK and Lee AH (2001) Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine* **20**, 2907–20.

HODDER

JOURNAL……ST **ARTID........ST084OA**

TO: CORRESPONDING AUTHOR

AUTHOR QUERIES - TO BE ANSWERED BY THE AUTHOR

The following queries have arisen during the typesetting of your manuscript. Please answer these queries by marking the required corrections at the appropriate point in the text.

Q1	Please check the “deletion of K” is correct.	
Q2	Please check “nijo” is correct.	
Q3	Please provide page range for “Bohning et al., 1997” reference.	
Q4	The reference Jiang (1998) is not cited in the text. Please check.	